Methods 151 (2018) 12–20

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Bi-clustering of metabolic data using matrix factorization tools

Quan Gu [a], Kirill Veselkov [b,*]

[a] MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Garscube Estate, Glasgow G61 1QH, UK
[b] Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, Exhibition Road, South Kensington, London SW7 2AZ, UK

ARTICLE INFO

Article history:
Received 17 January 2018
Received in revised form 4 February 2018
Accepted 6 February 2018
Available online 10 February 2018

Keywords:
Bi-clustering
Matrix factorization
Bi-cross validation
Metabolic data

ABSTRACT

Metabolic phenotyping technologies based on Nuclear Magnetic Spectroscopy (NMR) and Mass Spectrometry (MS) generate vast amounts of unrefined data from biological samples. Clustering strategies are frequently employed to provide insight into patterns of relationships between samples and metabolites. Here, we propose the use of a non-negative matrix factorization driven bi-clustering strategy for metabolic phenotyping data in order to discover subsets of interrelated metabolites that exhibit similar behaviour across subsets of samples. The proposed strategy incorporates bi-cross validation and statistical segmentation techniques to automatically determine the number and structure of bi-clusters. This alternative approach is in contrast to the widely used conventional clustering approaches that incorporate all molecular peaks for clustering in metabolic studies and require *a priori* specification of the number of clusters. We perform the comparative analysis of the proposed strategy with other bi-clustering approaches, which were developed in the context of genomics and transcriptomics research. We demonstrate the superior performance of the proposed bi-clustering strategy on both simulated (NMR) and real (MS) bacterial metabolic data.

1. Introduction

Modern Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry (MS) technologies generate vast amounts of unrefined metabolic data in biomedical studies [1,2]. The metabolic signature of a complex biological mixture ('metabolic profile'), such as that obtained from analysis of biofluids, consists of overlapping signals of hundreds to thousands of distinct chemical entities influenced by genes, treatment, gut microbiota and other environmental factors. This myriad of factorial influences results in complex inter-relationships between both spectral observations and variables. The clustering and related unsupervised learning tools are frequently used to discover patterns of relationships between samples and metabolites [3,4].

Given a two-dimensional data matrix X with m rows (samples) and n columns (variables), traditional clustering analysis aims to identify groups of samples (or respectively variables) that exhibit similar behaviour across all variables (or respectively samples). This strategy is useful to perform global partitioning of the data matrix. In "-omics" studies, molecules (e.g., genes or metabolites) can be involved in one or more biological processes and exhibit similar patterns of behaviour across a subset of samples (but not necessarily all). The bi-clustering strategies are more suitable in such cases. The objective of biclustering is to perform simultaneous clustering of both rows and columns in the data matrix [5]. This means that clustering derives a global model, while biclustering produces a local model. Each row in a bicluster is selected using only a subset of the columns and each column in a bicluster is selected using only a subset of the rows.

In "omics" sciences, the (bi)clustering methods have been widely applied to gene expression data matrix, where rows represent gene transcripts and column represent conditions/samples. The data matrix element corresponds to the expression level of a gene under a specific condition [6]. Unlike clustering algorithms, the goal of the technique is to identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions. Such biclusters are biologically relevant since they not only capture the correlated genes but also identify the genes that do not behave similarly in all conditions [7]. Thus, the biclustering algorithms have been shown to discover more biologically relevant clusters, compared to conventional global clustering techniques.

The biological application of biclustering algorithm was first used by Cheng and Church (CC) [8] for gene expression data. After this initial approach, a number of biclustering algorithms including Spectral [9], Plaid [10], BiMax [11], Xmotifs [12], OPSM [13], ISA [14], QUBIC [15] and FABIA [16] have been developed to identify

* Corresponding author.
E-mail address: kirill.veselkov04@imperial.ac.uk (K. Veselkov).

https://doi.org/10.1016/j.ymeth.2018.02.004
1046-2023/

various types of biclusters for gene expression data. A useful criterion to evaluate a biclustering algorithm is based on the identification of the type of biclusters the algorithm is able to find. In broad terms, the types of biclusters can be divided into [5]:

1. Biclusters with constant values.
2. Biclusters with constant values on rows or columns.
3. Biclusters with coherent values (addictive model).
4. Biclusters with coherent values (multiplicative model).
5. Biclusters with coherent evolutions.

The spectroscopic data (NMR or MS), with rows representing the samples and columns representing variables, can be considered as a linear combination of metabolite peaks plus noise, which is corresponding to the biclusters with constant values on columns or overlapped. However, the aforementioned biclustering methods with low anti-disturbing and fault tolerance are only suitable for gene expression data; in this scenario, the rows (genes) and columns (conditions) are fixed without overlapping partitions of conditions from the experiment.

Matrix factorization is a decomposition of a data matrix into a product of matrices, either for regularization or for interpretation. A variety of matrix factorization methods by incorporating different constraints, e.g. singular value decomposition (SVD) [17], principal component analysis (PCA) [18], and non-negative matrix factorization (NMF) [19], could be applied to minimise the dimensionality of the data yielding a representation of conditions as a linear combination of a reduced set of factors [20]. The factor scores/loadings represent sets of rows or columns that behave in a strongly correlated manner with the original data. In the gene expression data, various matrix factorization methods have been used to cluster genes or conditions based on local patterns and predict functional relationships [21,22]. Apart from the genomic datasets, the matrix factorization tools could also be used for exploring the spectroscopic datasets for their output matrices represent the relevance of samples and compound variables simultaneously [23].

In this paper, we first present a biclustering technique based on matrix factorization to identify subsets of correlated metabolites exhibiting similar patterns of behaviour across a subset of samples (but not necessarily all). The critical factor is how to select the number of biclusters. We have thus incorporated the bi-cross validation and statistical segmentation techniques to automatically determine the number and structure of bi-clusters. This alternative approach is in contrast to the widely used conventional clustering approaches that use all molecular peaks for clustering in metabolic studies and require *a priori* specification of the number of clusters. We perform the comparative analysis of the proposed strategy with other bi-clustering approaches, which were developed in the context of genomics and transcriptomics research.

2. Materials and methods

2.1. Synthetic dataset

A bicluster is a subset of rows that exhibit similar behaviour across a subset of columns, and vice versa. Given a data matrix A, a bicluster $A_{IJ}(I,J)$ denotes the submatrix of A that contains only the elements a_{ij} belonging to the submatrix with a set of rows I and set of columns J.

There are two ways to generate synthetic NMR/MS datasets for evaluating the algorithm performance of the metabolites biclustering. Considering the characteristic of NMR/MS spectroscopic data matrix, the first way is to use biclusters with constant values on columns to simulate the compound. Given the K bicluster, the dataset of i rows and j columns can be built by the equation

$$X_{ij} = \mu + \sum_{k=1}^{K} \beta_{jk}\rho_{ij}\kappa_{ij} + \varepsilon_{ij} \quad (1)$$

where X_{ij} represents the element in the bicluster, μ is the typical value and β_{jk} is the value for column j within the within the k-th bicluster. ρ and κ are indicator variables for row i and column j membership in the bicluster k. The noise ε is generated from a Gaussian distribution with zero mean and a varying standard deviation, i.e. $\varepsilon \in N(0,\delta)$, where δ is the noise level. Fig. 1 provides a simple example of 8×8 matrix with $\beta = \{1,2,3\}$. The typical value μ is set to 0 and the noise level δ is set to 0.1.

Aside from the method mentioned above, the MetAssimulo [24] is an important tool to simulate ^{1}H NMR spectra of metabolic profiles. MetAssimulo is a package, which can create realistic metabolic profiles containing large numbers of metabolites with a range of user-defined properties based on the concentration information input by the user or constructed automatically from the Human Metabolome Database. For instance, if the concentration information is in the custom mode, the user could set the concentration information of 'case' samples by defining the fold-change of the mean and standard deviation of corresponding concentration in 'control' samples, which is constructed automatically from the Human Metabolome Database. Furthermore, MetAssimulo is able to simulate shifts in NMR peak positions that result from matrix effects (e.g., pH variation), which are often observed in metabolic NMR spectra.

-0.04	-0.06	-0.12	0.07	0.02	0.02	-0.07	0.01
0.06	**1**+0.02	**2**-0.07	**3**+0.01	-0.03	0.04	-0.17	0.04
-0.01	**1**-0.09	**2**-0.06	**3**+0.05	-0.01	-0.06	0.09	0.01
-0.20	**1**-0.10	**2**+0.04	**3**+0.04	0.05	0.02	0.08	0.27
-0.09	-0.06	0.09	**2**-0.09	**3**+0.10	**1**+0.06	-0.08	-0.11
0.06	0.18	0.03	**2**+0.07	**3**-0.02	**1**+0.01	0.08	-0.18
-0.01	-0.1	-0.03	**2**-0.06	**3**+0.03	**1**+0.17	0.01	-0.11
-0.11	0.02	0.08	-0.06	-0.02	-0.05	0.02	-0.11

Fig. 1. A simple example of simulation data matrix with biclusters of constant values on columns.

In detail, the schema of building ^{1}H NMR synthetic data by MetAssimulo is listed as follows:

Step1: Given the number of biclusters K $\subset$ {1,2,. . .,k}, the number of samples in each biclusters M $\subset$ {m_1,m_2,. . .,m_k}, the number of metabolites in each biclusters N $\subset$ {n_1,n_2,. . .,n_k}.

Step2: For $k \in K$, $m \in M$ and $n \in N$: set the fold-change of mean and standard deviation of the concentration of n_k –th metabolite in m_k 'case' samples.

Step3: Run the MetAssimulo package and simulate ^{1}H NMR spectra of metabolic profiles, get the data matrix A and the median of the data matrix $A_{1/2}$.

Step4: Align and normalize the data matrix.

Step5: Logarithm transfers the data matrix $A' = \left|\frac{A}{A_{1/2}}\right|$ for biclustering analysis.

2.1.1. Non-negative matrix factorization

Non-negative matrix factorization, also known as the classical NMF model is a useful algorithm in multivariate analysis and linear algebra, which has been successfully applied in chemometrics [19]. The technique can be applied to the analysis of multidimensional datasets to reduce the dimensionality, discover latent patterns and aid in the interpretation of the data.

The main difference between NMF and other classical factorization techniques such as SVD [17] and PCA [18] depends on the non-negativity constraints imposed on both score and loading vectors. In this way, output matrices can be interpreted as parts of the data or as subsets of elements that tend to occur together in subportions of the dataset [20]. Thus, the factor matrices produced by NMF (i.e., factor scores and factor loadings) that lend themselves to a relatively easy contextual interpretation, while the factor matrices obtained by the other classical factorization approaches, allow themselves to be of the arbitrary sign with no obvious contextual meaning.

The NMF algorithm is described as follows:

$$A_{m\times n} = W_{m\times k}H_{k\times n} = \sum_{a=1}^{k} W_{m\times a}H_{a\times n} \tag{2}$$

where A is the positive data matrix with m samples and n variables, k is the number of components with $k << \min(m,n)$.

A solution to the NMF problem can be obtained by solving the following optimization object function:

$$\min_{W,H} f(W,H) = \frac{1}{2}\|A - WH\|_F^2 \tag{3}$$

where W is a basis matrix, H is a coefficient matrix, $\|\cdot\|_F$ is the Frobenius norm and W, $H \geq 0$ means that all elements of W and H are non-negative.

Considering the non-negativity of metabolites concentration in NMR/MS spectra, we use the NMF method to find biclusters of metabolites from NMR/MS spectra. As shown in the Eq. (2), under the number of biclusters k, the classical NMF approximately reproduce a ^{1}H NMR spectroscopic data matrix A of dimension m samples and n variables as a product of two non-negative constraint matrices W and H. The W factor scores matrix has the dimension of a single array (m samples) and k biclusters, while the columns of factor loading matrix H are known as variable vectors and are in one-to-one correspondence with the NMR/MS spectra data matrix A.

2.1.2. Bi-cross-validation

Given a large dataset matrix A of dimension $m \times n$, several useful methods handle it to produce two matrices W and H, and the cross-validation(CV) is a practical algorithm to determine the number of rank of W and H, i.e. the number of component in the large matrix [17]. However, the result is affected by the noise of matrix and there is a risk of overfitting. Considering the noise and complexity in spectroscopic data and the overlap between metabolites, the cross-validation is not suitable for predicting the number of biclusters in these data matrices.

As illustrated by Owen and Perry [17], bi-cross-validation algorithm (BCV) is a useful tool that is generally applicable to outer product approximations, just as CV is for independent and identically distributed random variables sampling. The performance and robustness of BCV is better than CV, and is more suitable than CV for the unsupervised learning (e.g. matrix factorization).

In the present study, we use BCV of matrix factorization to predict the number of the biclusters. The schema of algorithm is listed as follows:

Step1: Given a data matrix $A \in [0,\infty)^{m\times n}$, row and column holdout subset I_l = {1,2, ,m}, J_l = {1,2,. . .,n}, for number of holdout l = 1,2,. . .,L, and list of ranks the number of metabolites in each biclusters K = {1,2,. . .,min(m,n)}.

Step2: For $k \in K$:$BCV(k) \leftarrow 0$

Step3: For $l \in \{1,2,\ldots,L\}$ and $k \in K$:$I \leftarrow I_l$, and $J \leftarrow J_l$, fit the matrix factorization model: $A_{-I,-J} \doteq W^{(k)}_{-I,-J}H^{(k)}_{-I,-J}$.

Step4: Reconstruct the matrix and get the confirming residual matrix $\left\|A_{IJ} - A^{*(k)}_{IJ}\right\|_F^2$, where $A^{*(k)}_{IJ} \leftarrow A_{I,-J}(H^{(k)}_{-I,-J}W^{(k)}_{-I,-J})^{+}A_{-IJ}$.

Step5: Update the $BCV(k) \leftarrow BCV(k) + \left\|A_{IJ} - A^{*(k)}_{IJ}\right\|_F^2$.

2.1.3. Other methodologies

Apart from matrix factorization, a myriad of bicluster techniques have been proposed for gene expression data [5,11]. In this paper, the spectroscopic data (NMR or MS), rows represent the samples while columns represent variables, can be considered as a linear combination of metabolite peaks plus noise, which is corresponding to the biclusters with constant values on columns. For this reason, we compared our methods with the other biclustering techniques (e.g. Spectral [9], Plaid [10], BiMax [11], Xmotifs [12], ISA [14] and FABIA [16]) aimed at identifying constant columns.

Among these methods, the algorithms required for rescaling or iteration (e.g., Spectral and ISA) have longer running time on large datasets; the algorithms with methodology based on the binary value (e.g., BiMax, Xmotifs) are more sensitive to the noise of the data. Besides ISA, the performance of these methods is likely affected by the overlap of biclusters. Furthermore, Plaid, ISA and FABIA are also suitable for other bicluster classes. Brief methodological overview and the references of these biclustering techniques are listed in the Table 1.

2.1.4. Evaluating measurement

In this paper, we calculated the bicluster match score to evaluate the performance of biclsutering method, which was proposed by Prelic et al [11]. Without loss of generality, assume that s assigns larger scores to similar biclusters and smaller scores to dissimilar ones. M_1, M_2 are two sets of biclusters and the bicluster match score of sample S_s can be calculated by the following equation:

$$S_s(M_1,M_2) = \frac{1}{|M_1|}\sum_{b_1\in M_1}\max_{b_2\in M_2}\left(\frac{|b_1 \cap b_2|}{|b_1 \cup b_2|}\right) \tag{4}$$

where b_1, b_2 are biclusters in the set M_1, M_2 respectively, $|b_1 \cap b_2|$ is the number of data elements in their intersection, and $|b_1 \cup b_2|$ is the number in their union. Similarity, the bicluster match score of variable S_v can be calculated as well. The overall bicluster match score can be defined as $S(M_1,M_2) = \sqrt{S_s(M_1,M_2)\cdot S_v(M_1,M_2)}$.

Let E denote the ground truth bicluster set and F denote the set of found biclusters, the recovery score is $S(E,F)$ and relevance score is $S(F,E)$. If the recovery score S is maximized, it represents that the

Table 1
The summary of methods for identifying biclusters with constant row or column on the gene expression datasets.

Algorithm	Methodology	Description
BiMax [11]	Seeks the rectangles of '1''s in a binary matrix	Only suitable for the bicluster with constant up-regulated condition; sensitive to the noise and number of biclusters; affected by the overlap
Plaid [10]	Assume the bicluster is generated as the sum of a background effect, cluster effects, row effects, column effects and random noise	Both suitable for conditions of the bicluster with constant value and constant row/column; sensitive to the noise; affected by the overlap
Spectral [9]	Advantages over SVD spectral analysis of the original or rescaling raw data	Both suitable for conditions of the bicluster with constant up- or down-regulated condition; not sensitive to the noise; not suitable for the discrete datasets; limited in running speed on large datasets; affected by the overlap
Xmotifs [12]	A nondeterministic greedy algorithm that seeks biclusters with conserved rows/columns	Only suitable for conditions of the bicluster with constant row/column; required for dataset discretized and more sensitive to the noise; affected by the overlap limited in running speed on large datasets; affected by the overlap
FABIA [16]	Analysis for bicluster acquisition models the data matrix as the sum of biclusters plus additive noise, bicluster is the outer product of two sparse vectors	Both suitable for conditions of the bicluster with constant value and constant row/column; not sensitive to the noise and the number of biclusters; affected by the overlap
ISA [14]	A nondeterministic greedy algorithm that seeks biclusters from starting with a seed bicluster and re-running the iteration steps	Both suitable for conditions of the bicluster with constant value and constant row/column; not sensitive to the noise the number of biclusters, and the overlaps; limited in running speed on large datasets

algorithm is found in all the expected biclusters. Similarly, if the relevance score S is maximized, all the found biclusters were expected.

In this study, we develop a visualization graphical user interface and work on the spectroscopic dataset. The graphical user interface written in Matlab is available by contacting the corresponding author.

3. Results and discussion

3.1. Evaluation of biclustering on the synthetic dataset

We explored the bicluster model on the synthetic datasets. Firstly, we built a synthetic dataset with only 50 samples and 250 variables for observing the performance of biclustering meth-

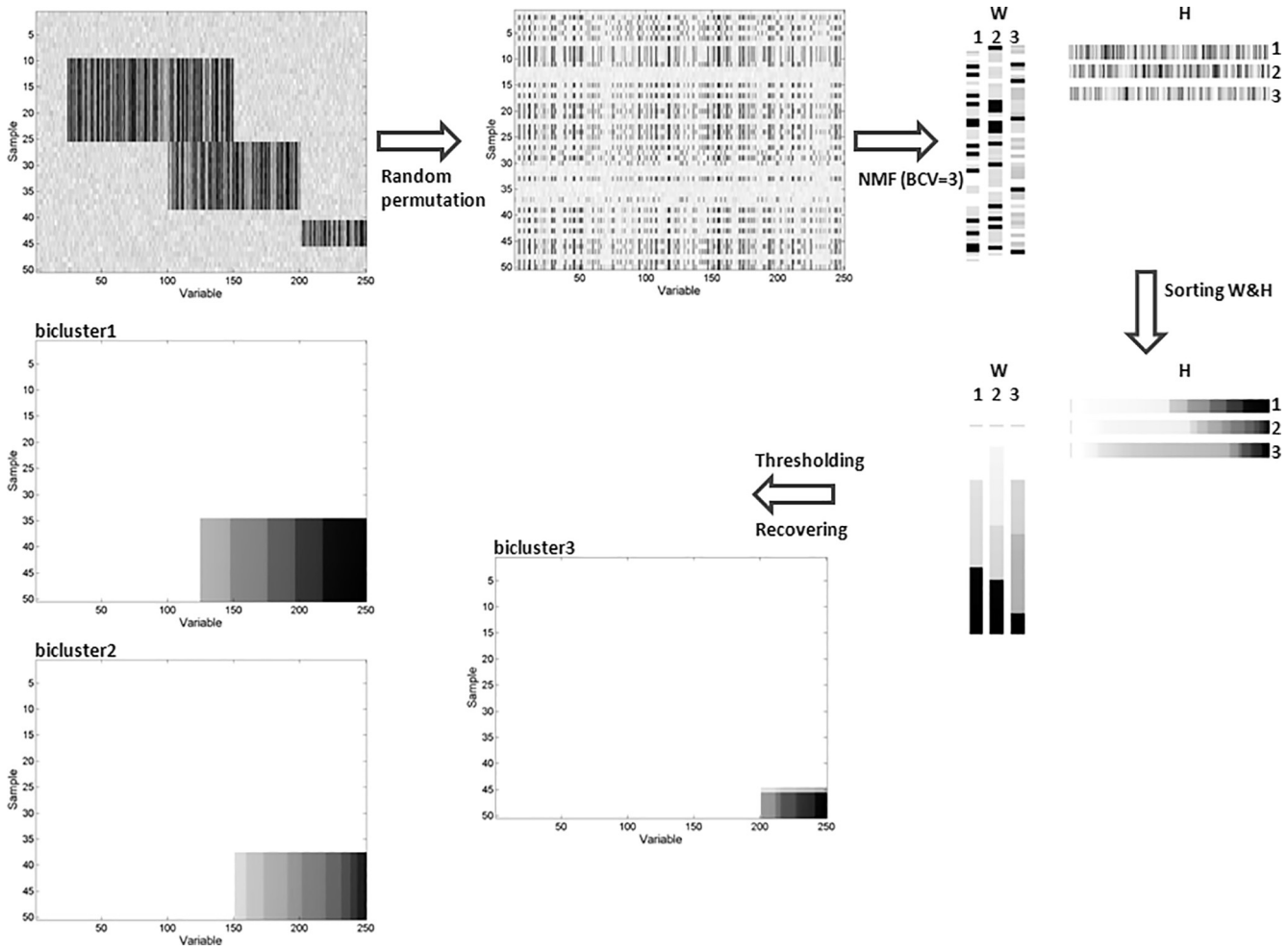

Fig. 2. General schema of the method NMF approximates the synthetic data 1(noise level $\delta = 0.25$) as a product of two submatrices, W and H. BCV is used for predicted the number of biclusters and thresholding algorithm is used for the identifying the indicator of each bicluster.

Table 2
Metabolites implanted of each bicluster in synthetic data 2 (the metabolites have the negative fold-change are in bold).

ID of bicluster	Metabolites in the bicluster
1	Citric acid, Creatine, Succinic acid, Hippuric acid, Serine
2	Creatinine, Citric acid, Glycine, **Formic acid**, Trimethylamine, **Hippuric acid**
3	**Creatinine, Formic acid**, Taurine, Betaine, **Guanidoacetic acid**, Trimethylamine

ods. From the Eq. (1), we set the number of biclsuter $k = 3$ with two of them having overlap. The value in each bilcuster is set as $\beta \in \{1, 2, \ldots, 5\}$, the typical value is set as the $\mu = \min(|X_{ij}|)$ meanwhile the noise level is set as $\delta \in [0, 1]$. Considering the randomness of the realistic datasets from experiment, we randomly permuted the samples (rows) and variables (columns) of the dataset and generated the synthetic dataset 1 (Fig. 2). The heatmap of the original dataset before rows and columns randomly permutated with the noise level $\delta = 0.25$. After the permutation of the rows and columns of the synthetic dataset, we calculated the average bicluster match scores to evaluate the performance of different biclustering models. The scheme of NMF model is shown in Fig. 2.

We also observed the performance of the bicluster algorithm on a synthetic dataset (synthetic data 2) with 30 samples generated by MetAssimulo. MetAssimulo can create realistic metabolic profiles containing large numbers of metabolites with a range of user-defined properties based on the concentration information input by the user or constructed automatically from the Human Metabolome Database. In the present study, the overall number of metabolites is 48, the number of biclusters is set to 3, and the fold-change values of mean and standard deviation of the concentration of metabolites are either positive or negative. Table 2 provides the metabolites implanted of each bicluster. Importantly, considering the peak shift of the output of MetAssimulo, the spectroscopic data is required for alignment and normalization. Following the recursive segment-wise peak alignment model [18] and logarithm transfer, we generated the synthetic dataset 2. As shown in the Fig. S1 (supplement file), the NMF model selects the correct ID of samples and hippuric acid (positive fold change of mean NMR spectra) within the bicluster 1 from the chemical shifts ^{1}H ppm 7.50–7.90.

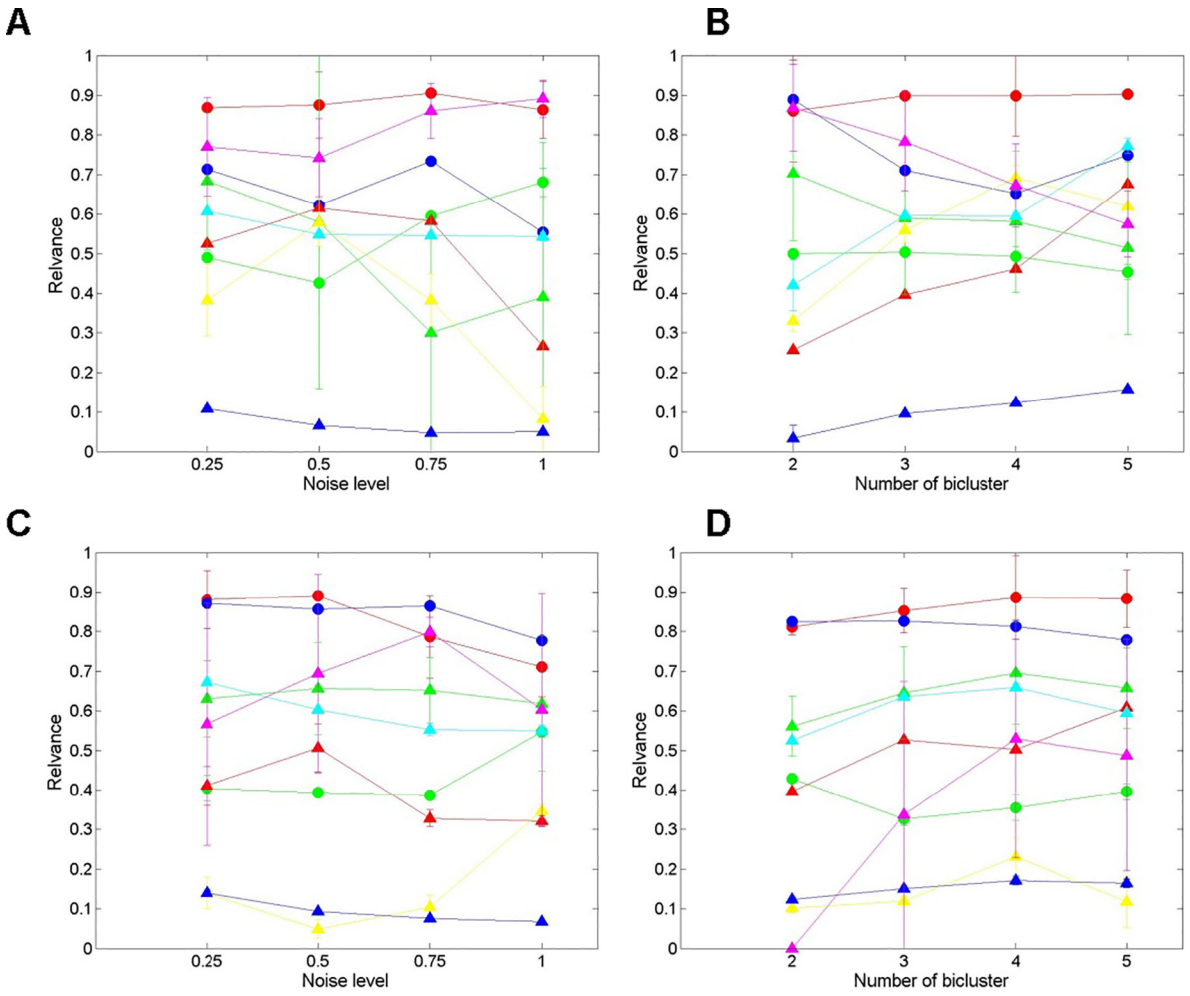

Fig. 3. The bicluster model experiment on the synthetic datasets. (A) The relevance score of bicluster methods on synthetic gene expression data [7] with different noise level; (B) The relevance score of bicluster methods on synthetic gene expression data [7] of the different number of biclusters; (C) The relevance score of bicluster methods on synthetic spectroscopic data 1 (Fig. 2) with different noise level; (D) The average relevance score of bicluster methods on synthetic spectroscopic data 1 (Fig. 2) the different number of biclusters. The methods: NMF (red circle), PCA (blue circle), Sparse NMF(green circle), BiMax(red triangle), Plaid(green triangle), Spectral(yellow triangle), Xmotifs (blue triangle), Fabia(magenta triangle) and ISA(cyan triangle. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Apart from our two synthetic datasets, we also used the simulation dataset generated for gene expression by Eren et al. [7] to validate the characteristics of the bicluster models. The matrix contains 500 genes (rows), 200 conditions (columns), and each bicluster with 50 rows and 50 without the overlap.

In this paper, we separately tested the performance of six classical biclustering models, i.e. Spectral [9], Plaid [10], BiMax [11], Xmotifs [12], ISA [14] and FABIA [16], and three matrix factorization algorithms, i.e. PCA, NMF and sparse NMF(SNMF) [25]. In matrix factorization algorithms, the number of biclusters could be predicted by BCV. Additionally, since the matrix factorization algorithms could only generate score and loading matrices, thresholding algorithms are utilized for selecting the samples and variables in each biclusters.

Fig. 3 summarises the relevance scores as a function of the number of biclusters and varying noise levels of the simulated datasets. We only represent the average relevance scores comparison but not the recovery score due to the similarity of both scores generated.

As expected, compared with six classical biclustering models, the matrix factorization methods have better robustness on the number of biclusters. It also validates the effectiveness of BCV prediction on the number of biclusters, which is an essential prerequisite for high quality performance. We tested the performance of six classical biclustering models and three matrix factorization algorithms on the simulation dataset of gene expression. The results of the average relevance scores for each bicluster model with different noise level and the number of biclusters are separately shown in the Fig. 3A and B. NMF achieves the highest relevance scores among the models, which validates the effectiveness of NMF model to identify the biclusters on the gene expression data. The methods Xmotif has the poorest performance on this dataset. With regards to the other methods, the performance of Plaid and FABIA is challenging as the number increased (Fig. 3B), whilst ISA and Spectral are negatively affected by the noise (Fig. 3A). However, as shown in the Figures, the SNMF is more sensitive to the noise than NMF. The reason is that the biclusters feature selection in sparse NMF is not based on the thresholding algorithm but on

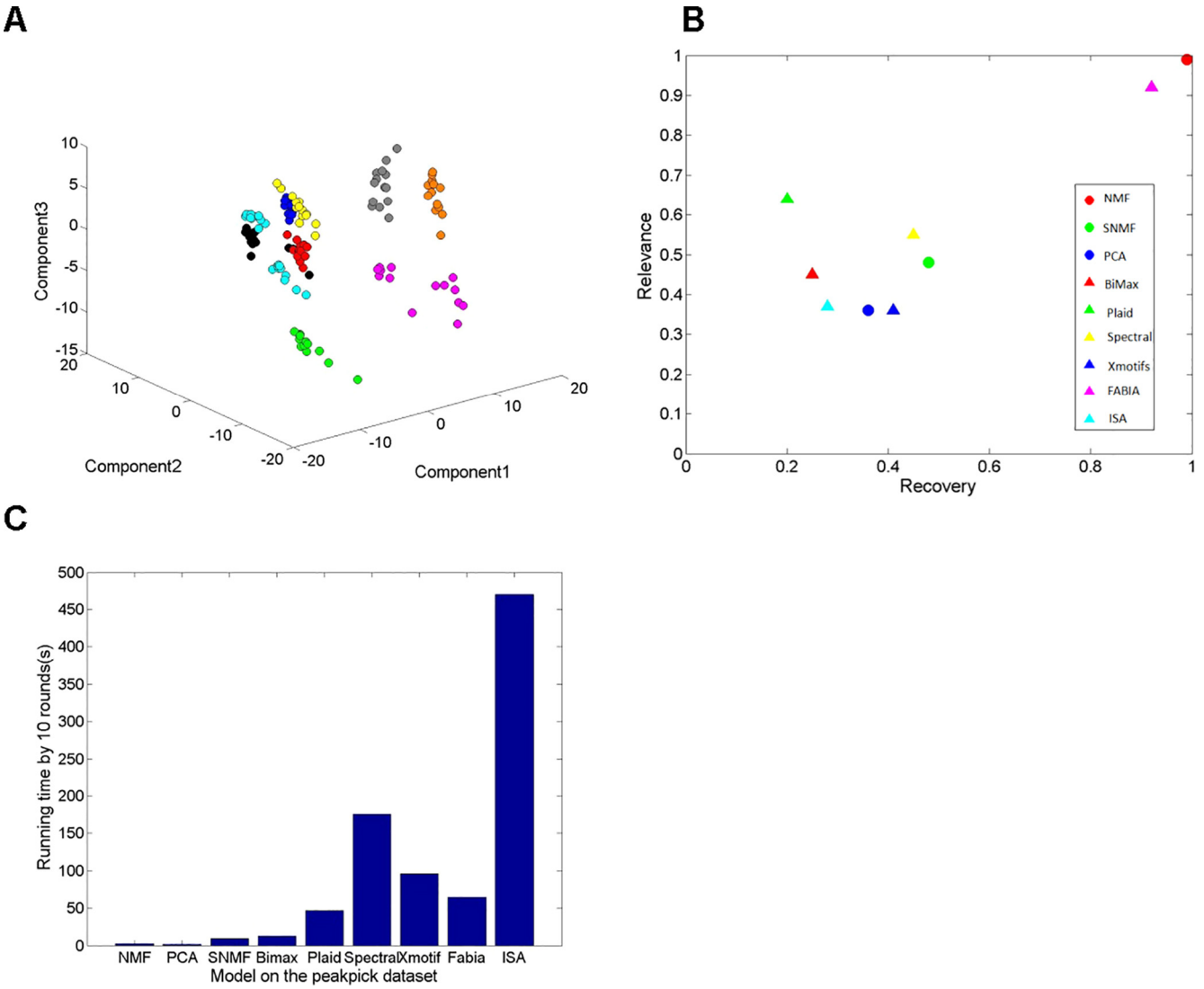

Fig. 4. The bicluster model experiment on the bacteria MS spectra[29]. (A) The 3-D PCA score plot of dataset. The species: *C. koseri* (red), *K. pneumonia* (green), *P. mirabills* (blue), *S. aureus* (magenta), *S. pyogenes* (yellow), *E. coli* (cyan), *P. aeruginosa* (black), *S. agalactiae* (brown), *S. marcescens* (orange); (B) The average recovery and relevance bicluster match score of different bicluster methods. The methods: NMF (red circle), PCA (blue circle), Sparse NMF(green circle), BiMax(red triangle), Plaid(green triangle), Spectral(yellow triangle), Xmotifs(blue triangle), Fabia(magenta triangle) and ISA(cyan triangle); (C) The running time of the algorithms running by 10 rounds on the peakpick dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the sparseness of loading or score values, which is not stable in case of the overlap of the biclusters.

On the synthetic dataset 1 of spectroscopic data, we tested the performance of the present nine algorithms as well. The plots of the average relevance scores for each bicluster model with different noise level and the number of biclusters are separately given in the Fig. 3C and D.

As indicated in the Fig. 3C and D, the overall bicluster match scores of NMF and PCA are obviously highest among all the models. It indicates that the inherent clustering property of matrix factorization is more suitable for the spectroscopic data. As shown in the plot, the algorithms that seek local patterns (e.g., Xmotifs) are more sensitive to the noise, while the algorithms which fit a model of the entire dataset (e.g., ISA, Plaid) are less sensitive. Bimax, which is required for pre-processing of the data matrix into binary, has poor performance although it is effective on the gene expression synthetic datasets. In general, NMF has the best performance among all the algorithms on both synthetic datasets.

3.2. Evaluation of biclustering on the biological dataset

We also explored the bicluster model on the MS dataset (negative ion) of nine bacterial species [29,30]. The clusters can be distinguished by the visualization of the first three principal components in the PCA (Fig. 4A) belonging to Gram-positive *Streptococcus spp.*, *Staphylococcus aureus*, Gram-negative *Pseudomonas aeruginosa* and a group consisting of five species that are not separated from each other and that all belong to the Enterobacteriaceae family (*Escherichia coli*, *Citrobacter koseri*, *Klebsiella pneumoniae*, *Serratia marcescens*, and *Proteus mirabilis*). To observe the performance of different algorithms, the MS dataset consisted of 135 samples and 185,001 *m*/*z* ratios that can be replaced by a reduced dataset of peak-picked variables with 135 samples and 1964 variables. Fig. 5 provides the NMF biclustering on this peak-picked dataset. If we reconstruct the selected bicluster, we will identify the samples and variables that are both associated in this bicluster.

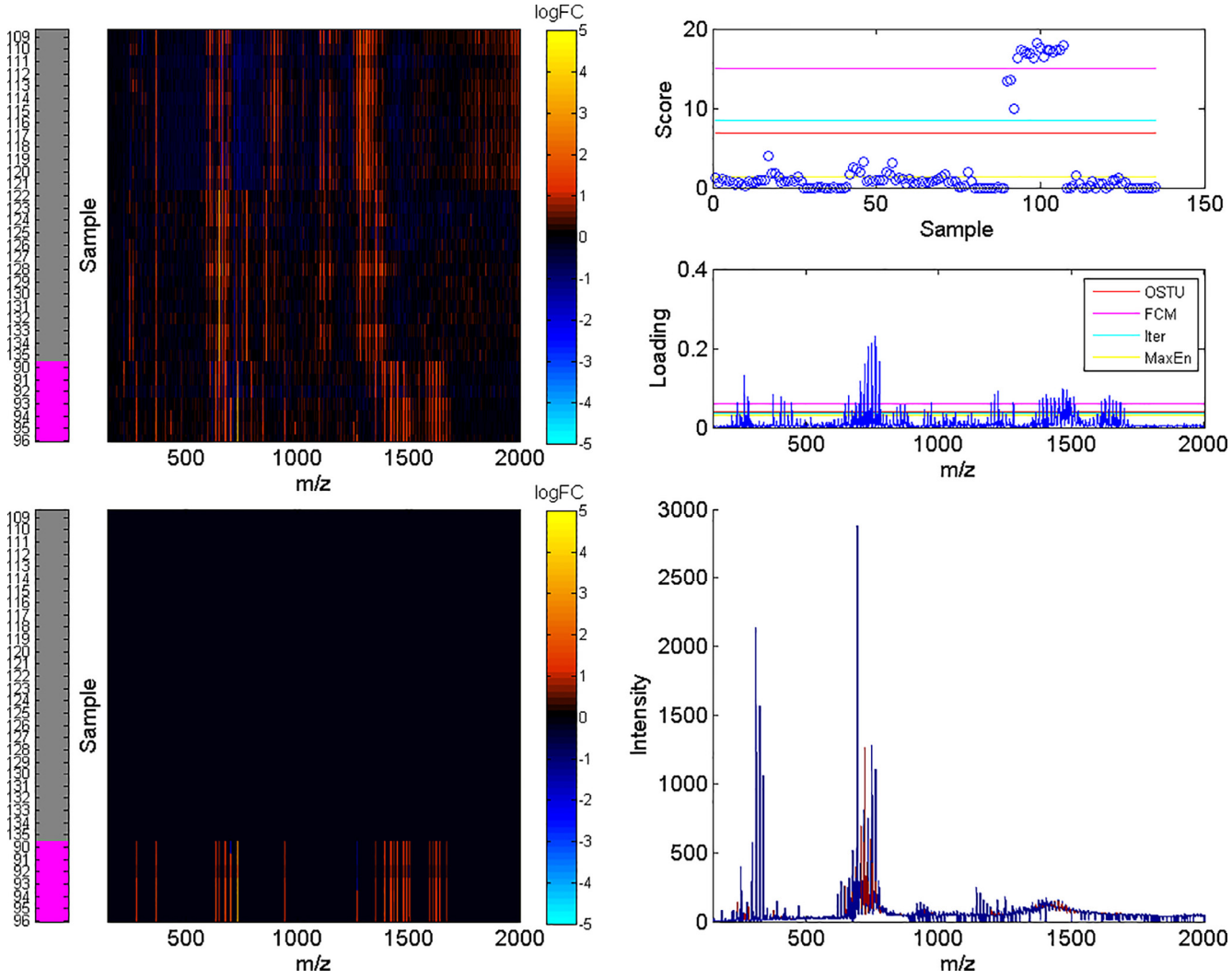

Fig. 5. The biclustering of bacteria MS peak-picked dataset using NMF: The right-bottom subplot shows the median of the MS spectra. The variables (right-bottom) associated with samples within the selected bicluster are separately marked with red. The left-top section of the plot indicates the log transform of fold change according to mean of the MS spectra. The left-bottom subplot represents the recovering of selected bicluster. The samples within the selected bicluster are marked with the magenta in the left colour bar. The right-top section of the plot represents the score and loading of the dataset and the comparison of thresholding algorithms: OTSU(red), FCM(magenta), Iter(cyan) and MaxEn(yellow). The horizontal axis (i.e. *m*/*z* ratio) of the subplot keeps in consistent with each other when selecting the compounds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
The bicluster match score of matrix factorization by thresholding methods on the synthetic data1 (noise level $\delta = 0.25$), synthetic data 2 and bacteria MS peak-pick dataset.

	PCA			NMF		
Thresholding	Synthetic data1	Synthetic data2	Bacteria MS data	Synthetic data1	Synthetic data2	Bacteria MS data
OTSU	0.72	0.61	0.36	0.91	0.79	0.99
FCM	0.65	0.53	0.44	0.62	0.61	0.87
Iter	0.72	0.61	0.28	0.85	0.77	0.99
MaxEn	0.57	0.59	0.23	0.61	0.60	0.47

Fig. 4B provides the overall prediction result on this MS dataset. It indicates that NMF achieves the highest bicluster match score on the metabolic data among the various methods. Moreover, NMF has the fastest speed among the various algorithms (Fig. 4C).

With a special focus on the thresholding algorithms, OTSU [26], fuzzy c-means (FCM) [27], iterative selection (Iter) [28] and maxime entropy (MaxEn) [26] are utilized for finding the bicluster from the score and loading generated by the matrix factorization methods. We compared the bicluster match score of these thresholding methods on our synthetic datasets and the bacteria MS (peak-picking) dataset. As shown in Table 3, the OTSU and Iter perform better than FCM and MaxEn. The performance of OTSU/Iter thresholding is still higher despite sparseness on the spectroscopic data, which validates our method (NMF with thresholding) as superior in recognizing the bicluster than SNMF.

The biclustering method is more useful on the MS dataset with sharp peaks and big signal-to-noise ratio than NMR data. Here, the bacterial dataset contains hundreds of unique spectral features with the signal to noise ratios (SNR) in the order of 10,000 times. High correlation between discriminating signals in each member of the same group. Moreover, the bacterial dataset has a hierarchical structure with the difference between Gram positive and Gram-negative bacteria (family level), representing where most the variance in the global data lies. With NMF (as with k-means), the number of output vectors (factors) is predefined, so the hierarchical nature is bypassed.

4. Conclusions

Spectroscopic data commonly contains around a thousand peaks from possibly hundreds of metabolites, is widely used in metabolomics to provide information on metabolite profiles of complex biological mixtures. This work represents a matrix factorization based biclustering model of spectroscopic data. In this paper, we use a novel bi-cross validation to decide the number of factors in the spectroscopic data matrix factorization tools. The simple thresholding methods (e.g., NMF) are used to transfer the spectroscopic data into two matrices. In this paper, we make the comparison among the various bi-clustering schemes on the simulation dataset, and the conclusion is that the simple matrix factorization tool is superior. Moreover, we develop a visualization graphical user interface and work on the bacterial dataset examples to test our biclustering model. The results demonstrate that the proposed matrix factorization based method for biclustering is useful for spectroscopic data. The future work would include the application of the proposed biclustering method on other biological datasets such as gene expression datasets.

Acknowledgments

We acknowledge the financial support for bioinformatics developments as part of MRC (MC_UU_12014/12), BBSRC (BB/L020858/1) and EU-METASPACE (34402) projects; KV acknowledges Waters corporation for funding and support throughout this study. Dr Sacheen Kumar (Division of Surgery, Imperial College London) is acknowledged for his advice and discussion. We acknowledge Prof Zoltan Takats from the Division of Computational and Systems Medicine, Imperial College London for the previous release and publication of the bacterial datasets as part of the above referenced BBSRC project [30]. Dr Tim Ebbels from the Division of Computational and Systems Medicine, Imperial College London is acknowledged for giving an access to the MetAssimulo package for the simulation of NMR datasets of complex chemical mixture.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ymeth.2018.02.004.

References

[1] M.-E. Dumas, S.P. Wilder, M.-T. Bihoreau, R.H. Barton, J.F. Fearnside, et al., Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models, Nat. Genet. 39 (2007) 666–672.
[2] J.K. Nicholson, E. Holmes, J.M. Kinross, A.W. Darzi, Z. Takats, et al., Metabolic phenotyping in clinical and surgical environments, Nature 491 (2012) 384–392.
[3] B.J. Blaise, L. Shintu, B. Elena, L. Emsley, M.-E. Dumas, et al., Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics, Anal. Chem. 81 (2009) 6242–6251.
[4] M.E. Dumas, A.R. Rothwell, L. Hoyles, T. Aranias, J. Chilloux, et al., Microbial-host co-metabolites are prodromal markers predicting phenotypic heterogeneity in behavior, obesity, and impaired glucose tolerance, Cell. Rep. 20 (2017) 136–148.
[5] S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey, Ieeeacm Trans. Comput. Biol. Bioinform. 1 (2004) 24–45.
[6] A. Tanay, R. Sharan, R. Shamir, Discovering statistically significant biclusters in gene expression data, Bioinformatics 18 (2002) S136–S144.
[7] K. Eren, M. Deveci, O. Küçüktunç, Ü.V. Çatalyürek, A comparative analysis of biclustering algorithms for gene expression data, Brief Bioinform. 14 (2013) 279–292.
[8] Y. Cheng, G.M. Church, Biclustering of expression data, Proc. Int. Conf. Intell. Syst. Mol. Biol. Ismb. Int. Conf. Intell. Syst. Mol. Biol. 8 (2000) 93–103.
[9] Y. Kluger, R. Basri, J.T. Chang, M. Gerstein, Spectral biclustering of microarray data: coclustering genes and conditions, Genome Res. 13 (2003) 703–716.
[10] H.L. Turner, T.C. Bailey, W.J. Krzanowski, C.A. Hemingway, Biclustering models for structured microarray data, Ieeeacm Trans. Comput. Biol. Bioinform. Ieee Acm. 2 (2005) 316–329.
[11] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, et al., A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics 22 (2006) 1122–1129.
[12] T.M. Murali, S. Kasif, Extracting conserved gene expression motifs from gene expression data, Pac. Symp. Biocomput. Pac. Symp. Biocomput. (2003) 77–88.
[13] A. Ben-Dor, B. Chor, R. Karp, Z. Yakhini, Discovering local structure in gene expression data: the order-preserving submatrix problem, J. Comput. Biol. J. Comput. Mol. Cell. Biol. 10 (2003) 373–384.
[14] S. Bergmann, J. Ihmels, N. Barkai, Iterative signature algorithm for the analysis of large-scale gene expression data, Phys. Rev. E Stat. Nonlin Soft Matter. Phys. 67 (2003) 031902.
[15] G. Li, Q. Ma, H. Tang, A.H. Paterson, Y. Xu, QUBIC: a qualitative biclustering algorithm for analyses of gene expression data, Nucleic Acids Res. 37 (2009) e101.
[16] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, et al., FABIA: factor analysis for bicluster acquisition, Bioinform. Oxf. Engl. 26 (2010) 1520–1527.
[17] A.B. Owen, P.O. Perry, Bi-cross-validation of the SVD and the nonnegative matrix factorization, Ann. Appl. Stat. 3 (2009) 564–594.
[18] K.A. Veselkov, J.C. Lindon, T.M.D. Ebbels, D. Crockford, V.V. Volynkin, et al., Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery, Anal. Chem. 81 (2009) 56–66.

[19] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[20] P. Carmona-Saez, R.D. Pascual-Marqui, F. Tirado, J.M. Carazo, A. Pascual-Montano, Biclustering of gene expression data by non-smooth non-negative matrix factorization, BMC Bioinf. 7 (2006) 78.
[21] J.-P. Brunet, P. Tamayo, T.R. Golub, J.P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization, Proc. Natl. Acad. Sci. USA 101 (2004) 4164–4169.
[22] P.M. Kim, B. Tidor, Subsystem identification through dimensionality reduction of large-scale gene expression data, Genome Res. 13 (2003) 1706–1718.
[23] J.K. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Metabonomics: a platform for studying drug toxicity and gene function, Nat. Rev. Drug Discov. 1 (2002) 153–161.
[24] H.J. Muncey, R. Jones, M.D. Iorio, T.M. Ebbels, MetAssimulo: simulation of realistic NMR metabolic profiles, BMC Bioinf. 11 (2010) 496.
[25] H. Kim, H. Park, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, Bioinformatics 23 (2007) 1495–1502.
[26] M. Sezgin, B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, J. Electron. Imaging 13 (2004) 146–168.
[27] D. Dembélé, P. Kastner, Fuzzy C-means method for clustering microarray data, Bioinformatics 19 (2003) 973–980.
[28] H.J. Trussell, Comments on "picture thresholding using an iterative selection method", Ieee Trans. Syst. Man. Cybern. 9 (1979) 311.
[29] N. Strittmatter, E.A. Jones, K.A. Veselkov, M. Rebec, J.G. Bundy, et al., Analysis of intact bacteria using rapid evaporative ionisation mass spectrometry, Chem. Commun. 49 (2013) 6188–6190.
[30] D. Galea, P. Inglese, L. Cammack, N. Strittmatter, M. Rebec, et al., Translational utility of a hierarchical classification strategy in biomolecular data analytics, Sci. Rep. 7 (2017) 14981.